The Report committee for Sandra Young Jackson

Certifies that this is the approved version of the following report:

Utilizing Socio-Economic Factors to Evaluate Recruiting Potential for a US Army Recruiting Company

APPROVED BY

SUPERVISING COMMITTEE:

_____

Nedialko Dimitrov, Supervisor

_____

Jonathan K. Alt

Utilizing Socio-Economic Factors to Evaluate Recruiting Potential for a US Army Recruiting Company

by

Sandra Young Jackson, B.S.

Presented to the Faculty of the Graduate School

of The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

May 2015

Acknowledgments

Thank you to CRJ, ZFJ, and ZRJ for ZFJ & ZRJ, helping me win, and not letting me procrastinate respectively. Thank you to Dr. Ned Dimitrov my advisor, LTC Jonathan K. Alt, my second reader, and my professors for your time and teaching.

Utilizing Socio-Economic Factors to Evaluate Recruiting Potential for a US Army Recruiting Company

Sandra Young Jackson, M.S.E.

The University of Texas at Austin, 2015

SUPERVISOR: Nedialko B. Dimitrov

In order to maintain military strength, the United States Army is consistently challenged with recruiting new soldiers. Currently the Army evaluates its recruiting capacity by calculating a weighted average of the previous four years of recruiting data. This report provides: (1) a description of the current method of calculating recruiting capacity; (2) an alternative approach for the calculation; and (3) an evaluation process and corresponding results to identify effective recruiting capacity methods. Specifically, the study analyzes the effectiveness of multiple linear regression and Poisson regression models to compute recruiting capacity. Surprisingly, even though essentially all previous literature on recruiting suggests Poisson regression to model recruiting arrival rates, we show strong empirical evidence that multi-linear regression is a better modeling tool than Poisson regression for the recruiting data. On out-of-sample tests involving 32 competing models, the negative log-likelihood for the multi-linear regression models is, on average over all the models, 11% smaller than the corresponding Poisson regression model. On out-of-sample tests involving an additional 20 models, the negative log-likelihood for the multi-linear regression is on average 85% smaller than the corresponding Poisson regression. The statistical models for recruiter rate suggest there is great potential for recruiting capacity because socio-economic factors do not limit the number of recruits. In other words, the results suggest that if the Army wants to increase recruits, one additional recruiter results in an additional 0.89 recruits. Analysis of the explanatory power of different socio-economic factors identifies the population of

qualified military aged persons as a key indicator, followed by unemployment rate; however, further study is required to compile and evaluate additional socio-economic factors and their contribution to predicting numbers of recruits or the number of recruits per recruiter.

Table of Contents

Chapter 1 Introduction

Chief of Staff of the Army, 4-star General Raymond Odierno, said, "The strength of our nation is our Army, the strength of the Army is our soldiers." (USAREC Talking Points, 2013). The task of filling the Army's ranks belongs to US Army Recruiting Command (USAREC). USAREC, led by a 2-star General, is organized into six subordinate brigades with each brigade commanding up to eight subordinate battalions. Each battalion commands the subordinate companies in its area and each company commands specific recruiting stations. USAREC is staffed by over 9,500 soldiers and civilians with more than 1,400 recruiting stations throughout America and overseas. See Appendix A for an organizational chart (USAREC *About Us*, 2014). USAREC's explicit mission is to "provide the strength of the Army" (USAREC Manual 3-0, 2009).

A *recruit* is defined as a qualified civilian who signed a contract to serve the Army as a future soldier. The current method to determine the potential number of recruits is cumbersome and unwieldy. The main goal of this report is to examine a simpler approach to setting realistic recruiting goals. Such an approach would enable effective communication within USAREC, and potentially increase the accuracy of the Army's recruiting goals.

1.1 Problem Description

Currently, the Army computes the number of potential recruits through a three-step process that is essentially a weighted average of recruiting data from the last four years (Clingan & Stokan, 2009). The process is as follows:

Step 1: Partition zip codes by demography. Each zip code is divided by the demographic characteristics of race, age, and gender. Partitions are called *tactical segments*. Each tactical segment is standard across all zip codes. For example, tactical segment 4 in all zip codes, consists of the Caucasian males, age 16-19.

Step 2: Calculate the *best penetration rate*. The penetration rate is a weighted average of the fraction of the population recruited over the last four years. This weighted

average is calculated at the station and company level. The best penetration rate for a specific station is the maximum of the company and station calculations.

Step 3: Determine the *Army volume potential* (AVP) for each zip code. Utilizing the penetration rate found in Step 2, calculate the number of potential recruits in each tactical segment in a particular zip code. Sum all tactical segments in a zip code to arrive at the AVP for that zip code.

A weighted four year average of previous recruits sets the minimum goal for recruiters in the same zip code and tactical segment. The AVP sets the upper limit for what the Army considers reasonable to achieve in each zip code. See Appendix A for a detailed explanation of the recruiting capacity calculation.

Currently, the Army uses 39 different tactical segments to describe each zip code. However, it is soon transitioning to something call PRIZM Segments. PRIZM is a marketing tool created by Nielsen, a marketing information resources company. There are 66 different PRIZM segments. The calculations would essentially remain the same except that PRIZM segments will replace tactical segments (Stokan, 2014).

1.2 Report Goals and Methods

USAREC Manual 3-0 (2009) "acknowledges that socio-economic factors effect recruiting." This report explores these effects with the following goals:

1. Create a simple and accurate prediction method for recruiting capacity.
2. Identify the most significant socio-economic factors in determining recruiting capacity.

To achieve these goals, the report builds a predictive method for estimating recruiting capacity based on socio-economic factors. In fact, the report examines two such models, a multi-linear regression model and a Poisson regression model.

Chapter 2. Related Work

DoD transitioned to an all-volunteer force in 1973, heightening the need to study recruitment. Past studies of recruiting fall into two main categories, social-science discourse and mathematically based models. Studies rooted in social-science highlight the importance of socio-economic factors while studies employing mathematical modeling produce quantifiable results. In this chapter we compare and contrast our study to this past literature. In addition, the related works serve as a method to choose socio-economic factors in a deliberate rather than ad-hoc manner. Finally, the chapter also highlights a few other documents important to Army recruiting, including Army documentation dictating the actual execution of the recruiting mission.

2.1 Social Science Approach

Recruiting for military service is an art in which a recruiter must appeal to an individual's motivations and desires. This study seeks to account for the factors affecting potential recruits. Social-science based works offer excellent insight on which key variables to use as input in the predictive model.

Stephen Foti (1978) studied the impact of socio-economic factors on recruiting from an operations management point of view. He noted that rising unemployment and an uncertain economy helped all services achieve recruiting goals immediately after the transition to an all-volunteer force. Quester (2005) examined changes in the demography of the United States and postulated the consequences for the military. He also noted that unemployment and a downturned economy improved recruitment. He ultimately asserted that the Army needs to account for demographic trends because they exert pressure on the recruiting mission.

Social-science based studies seek to draw general conclusions from big-picture concepts of human behavior. This study differs from social-science based works because it seeks to utilize quantifiable socio-economic factors in a mathematical model to predict the number of recruits at the company level.

2.2 Mathematical Modeling Approach

Of works utilizing a mathematical model, there are two kinds. The first kind applies mathematical optimization techniques to place recruiting stations and determine recruiter strength (Schwartz, 1993). This study does not focus on that type of modeling. The second kind focuses on mathematical models to determine recruiting potential. This study is most closely associated with these recruiting potential studies, which focus on accounting for socio-economic or marketing factors in the recruiting problem.

One study, conducted by the RAND Corporation, specifically examines the marketing tools utilized in order to meet the recruiting mission for all four military services in the fiscal year 1997 (McDonald & Murray, 1999). In recent years, US Navy in particular has increased its efforts to apply mathematical modeling to the recruiting problem. These efforts include a study by Pinelis, Schimitz, Miller, and Rebhan. Pinelis et al (2011) studied the supply of eligible recruits as an allocation tool for the recruiting mission. They produce results at the Navy Recruiting District (NRD) level. An NRD is approximately equivalent to an Army battalion. Another naval study, conducted by Evans and Powell (2014), developed a metric called the Nobel Index. This Nobel Index rates the recruiting production of a specific geographic area. The most recent naval study, produced by Williams (2014), developed models to assess a Navy Recruiting Station's (NRS) recruiting potential. An NRS is approximately equivalent to an Army Recruiting Company.

These mathematical model based studies employ either multi-linear regression or some form of Poisson regression. The primary difference between all studies are the specific socio-economic factors utilized to explain the dependent variable. A key difference between this report and previous studies is how the regression models are developed. This study goes further than previous studies, by assessing the accuracy of the models using *out of sample cross validation* tests. Out of sample cross validation is defined in Section 3.4. In other words, this study measures predictive accuracy of the models by running predictions on data the model has never seen before. These measures

of predictive accuracy allow this study to characterize which mathematical models are, in fact, the better performers.

2.3 Army Documents

The final set of related works are Army produced documents. There are 45 Army regulations, six pamphlets, two supplements, and six manuals governing recruiting operations (USAREC *Electronic Publications*, 2014). While none are academic studies, they inform this study of Army motivations and current methods. In addition to these published documents, there are other supporting documents from USAREC's intelligence section specifically detailing how the Army currently calculates recruiting potential (Clingan & Stokan, 2009).

Chapter 3. Data and Methodology

This chapter reviews the data used to construct the models and explains the methods behind the multi-linear regression and Poisson regression models. The data set contains 10,323 observations which encompasses four fiscal years' worth of recruiting data from 2011-2014. It only contains active recruiting companies so if a company existed in 2011 but not 2013, it is entirely absent from the data. Included in this study are 250 recruiting companies.

3.1 Input Data

USAREC provided the source data for this study. It is organized into one file and aggregated by recruiting company. In other words, each row specifies the data for a specific recruiting company's catchment area. Table 1 shows a sample of the input data:

RSID	QMA	Year	Month	Recruits	Unemployment Rate	Recruiters	Metro	Micro	Other
1A1	120763	2011	3	27	8.56	26	166	56	42
1A3	88824	2011	3	12	7.84	22	95	202	132
1A4	190495	2011	3	28	9.19	37	218	31	11

Table 1. Sample of the input data file. The total data file contains 10,323 observations. Each row is the historical recruiting performance of a specific company in a specific month, along with the socio-economic factors for the company's catchment area at the time.

The columns are defined as follows:

RSID: The unique identifier for an Army recruiting company.

Qualified Military Age (QMA): The population of men and women between 17 and 24 years old.

Year: The calendar year from which the data came.

Month: The calendar month from which the data came.

Recruits: The number of recruits achieved.

Unemployment Rate: The unemployment rate.

Recruiters: The number of recruiters assigned.

Metro: The number of zip codes with populations over 50,000.

Micro: The number of zip codes with populations between 10,000 and 50,000.

Other: The number of zip codes with populations less than 10,000.

3.2 Multi-linear Regression

A basic multi-linear regression (MLR) model has the form:

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i \, .$$

This means the dependent variable (y_i) is predicted by a regression coefficient (β_n) multiplied by an explanatory (also called independent) variable (x_{in}) plus an error term (ε_i). In this model, (x_{i0}) is set to 1 to provide a constant term. The assumptions of a multi-linear regression model are:

1. Errors are normally distributed.
2. Errors have constant variance.
3. Errors are independent and uncorrelated.
4. The model is structurally sound in that the dependent variables are able to be explained by a linear approximation of the explanatory variables.

There are a number of techniques used to determine goodness of fit for a multi-linear regression model. This report utilizes the negative log-likelihood statistics to determine goodness of fit as well as compare the results of this model against the Poisson regression results. Traditional R^2 analysis is included in Appendix C.

The indices and variables specific to this study are as follows:

Indices:

i: a specific company, in a specific month and year – we call this an observation

n: a specific socio-economic factor set as an explanatory variable – for example, unemployment rate

Variables:

y_i: number of recruits

x_{in}: value of explanatory variable

β_n: regression coefficient

ε_i: error term

This multi-linear regression model attempts to explain the number of recruits as a linear combination of various socio-economic factors and a measurement of the Army's effort – the number of recruiters. The study creates 16 different versions of the model in an effort to determine which socio-economic factor most strongly affects the number of recruits. See Table 2 for a list of the multi-linear regression models and the explanatory variables considered in each version.

Version	Explanatory Variable						
1	x_{i1} = Unemployment Rate						$\beta_0 x_{i0}$ = constant
2	x_{i1} = Metro						$\beta_0 x_{i0}$ = constant
3	x_{i1} = Micro						$\beta_0 x_{i0}$ = constant
4	x_{i1} = Other						$\beta_0 x_{i0}$ = constant
5	x_{i1} = QMA						$\beta_0 x_{i0}$ = constant
6	x_{i1} = Recruiter						$\beta_0 x_{i0}$ = constant
7	x_{i1} = Unemployment Rate		x_{i2} = Recruiter				$\beta_0 x_{i0}$ = constant
8	x_{i1} = Metro		x_{i2} = Recruiter				$\beta_0 x_{i0}$ = constant
9	x_{i1} = Micro		x_{i2} = Recruiter				$\beta_0 x_{i0}$ = constant
10	x_{i1} = Other		x_{i2} = Recruiter				$\beta_0 x_{i0}$ = constant
11	x_{i1} = QMA		x_{i2} = Recruiter				$\beta_0 x_{i0}$ = constant
12	x_{i1} = Unemployment Rate		x_{i2} = Metro		x_{i3} = Recruiter		$\beta_0 x_{i0}$ = constant
13	x_{i1} = Unemployment Rate		x_{i2} = Micro		x_{i3} = Recruiter		$\beta_0 x_{i0}$ = constant
14	x_{i1} = Unemployment Rate		x_{i2} = Other		x_{i3} = Recruiter		$\beta_0 x_{i0}$ = constant
15	x_{i1} = Unemployment Rate		x_{i2} = QMA		x_{i3} = Recruiter		$\beta_0 x_{i0}$ = constant
16	x_{i1} = Unemployment Rate	x_{i2} = Metro	x_{i3} = Micro	x_{i4} = Other	x_{i5} = QMA	x_{i6} = Recruiter	$\beta_0 x_{i0}$ = constant

Table 2. Model Versions and Explanatory Variables.

3.2.1 Multi-linear Regression Negative Log-Likelihood Statistic

This study builds each multi-linear regression model by minimizing the sum of squared errors. However, to compare multi-linear regression to Poisson regression, the study needs a negative log-likelihood statistic. Recall that one of the assumptions for a multi-linear regression is that the errors are normally distributed. This means the errors have a canonical *probability density function* (PDF). A logarithm is a continuous strictly increasing function over the range of the PDF. Values that maximize the PDF also

maximize a natural logarithm transformation of the PDF, also called a log-likelihood function. Let *M* annotate a specific model version, *D* be a set of observations, β_M^T be the transposed vector of regression coefficients, and x_M be the vector of explanatory variable values. Then the log-likelihood function utilized is:

$$\ln\big(L(\beta_M|X,Y)\big) = \sum_D [-\frac{1}{2}\ln(\sigma^2 2\pi) - \frac{(y_i - \beta_M^T x_M)^2}{2\sigma^2}].$$

where the variance of the errors, σ^2, is:

$$\sigma^2 = \sum_D \frac{(y_i - \beta_M^T x_M)^2}{|D|}.$$

Recall that *i* is a specific observation.

3.3 Poisson Regression

A Poisson regression is a form of predictive model that only produces values between zero and positive infinity and is particularly useful when modeling count data. In this type of regression, the model assumes the dependent variable has a Poisson distribution. The natural logarithm of the rate of the Poisson distribution is modeled as a linear combination of explanatory variables. The rate of a Poisson distribution is also its expected value. The basic form for a model with dependent variable *y* following a Poisson distribution with rate λ, where λ is the expected value of *y*, *E(Y)* is:

$$\ln(E(Y)) = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}\ .$$

This means that $\ln(E(Y))$ is predicted by a regression coefficient (β_n) multiplied by a explanatory variable (x_{in}) and (x_{i0}) is 1. More compactly:

$$\ln(E(Y)) = \beta_M^T x_M \, .$$

where β_M^T is the transposed vector of regression coefficients, x_M is the vector of explanatory variable values, and *M* denotes a specific model.

Assumptions for a Poisson regression model are:

1. The dependent variable *y* follows a Poisson distribution with rate λ where λ is *E(Y)*.
2. If the observations of y_i are independent with corresponding values of x_{in}, then B_M can be estimated by maximum likelihood. Specifically, the coefficient vector B_M is estimated by minimizing the negative log-likelihood (NLL):

$$l(\beta_M|X,Y) = -\sum_{i=1}^{n} \left(y_i \beta_M^T x_M - e^{\beta_M^T x_M} - \ln(y_i!) \right).$$

Minimizing the Poisson NLL is fundamentally different than a multi-linear regression on a natural log transformed dependent variable. In other words, to find the parameters, β_M, for a set of observations, we minimize the above function, as opposed to a sum of squared errors.

This study specifically focuses on predicting the number of recruits achievable by a recruiting company. This is directly in line with the concept of count data.

The indices and variables specific to this study are as follows:

Indices:

i: a specific observation

n: a specific socio-economic factor set as an explanatory variable – for example, unemployment rate

Variables:

y_i: observed number of recruits

x_{in}: value of explanatory variable

β_n: regression coefficient

As with the multi-linear regression model, 16 different model versions are created utilizing Poisson regression. The combination of explanatory variables used in Poisson regression is the same as those used in multi-linear regression. See Table 2 for a complete list of model versions and explanatory variable combinations.

This report utilizes the negative log-likelihood statistics to determine goodness of fit as well as compare the results of this model against the multi-linear regression results. *Psuedo-*R^2 analysis is defined and included in Appendix C.

3.4 Out of Sample Cross-validation

To prevent over-fitting the data in both multi-linear regression and Poisson regression, the study applies a process, call it *out of sample cross-validation*, for model selection and validation. The process proceeds in the following steps:

Step 0. Data and Model Preprocessing: Before beginning the cross-validation process, first organize the data and define the models.

1. Data Preprocessing: Randomly partition the data into *k* sets, call them $D_1, D_2, \ldots, D_k$. Hold the *k*th set to the side and call it the *final test data set*.
2. Model Preprocessing: Define the *model space*. The model space is the set of models under consideration, in this study, there 32 proposed models–16 for each regression type. Each of the 16 versions rely on a different set of explanatory variables.

Step 1. Model selection: Select the best model using maximum likelihood and out-of-sample cross validation on data sets D_1 through D_{k-1}.

1. Cross-validation is as follows: Using data sets D_1 through D_{k-1} iterate over *i*, where $i = 1 \ldots k-1$. Fit the model on the union of data sets $D_1, D_2, \ldots D_{i-1}, D_{i+1}, \ldots, D_{k-1}$. Call the result of fitting model *M* to this data M_{-i}. For example, if *i* = 3, fit the model on the union of the data sets $D_1, D_2, D_4, \ldots, D_{k-1}$. Call the resulting model M_{-3}.
2. Evaluate M_{-i} on data set D_i. The evaluation we use is the likelihood of observing D_i under M_{-i}. Call this likelihood L(M_{-i}, D_i).
3. We define the *average likelihood* of model *M* as $AL(M) = \frac{1}{k-1} \sum_i L(M_{-i}, D_i)$
4. We select model *M** that maximizes the average likelihood.

Step 2. Model validation: As a final step, validate model *M** is indeed a good fit for the data.

1. Train the model *M** on data set D_1 through D_{k-1}. Call this model M^*_{-k}.
2. Evaluate $L(M^*_{-k}, D_k)$ on the final test data set.

This study sets $k = 11$ and randly selects 1,323 observation of the total data set as final test data set, D_k. The remaining 9,000 are randomly partitioned into 10 sets, $D_1, D_2, \ldots, D_{10}$. These sets are used for cross-validation and to train the model for the final model validation step. The 10 bins of training data mean the study employs 10-fold cross-validation for model selection.

The out-of-sample cross validation method described above is constructed to evaluate and select models based on their predictive ability on data they have not fit. The model selection step, Step 1, selects models based on their average predictive ability on data they have not fit. Then, Step 2 Model validation, ensures that we have not somehow over-fit the data in Step 1 –having a large model space in Step 1 may over-fit data even though we are evaluating the models on out-of-sample predictive ability.

Chapter 4. Results

The study results are presented in phases. The Section 1 reviews the results using in-sample training data and gives model regression coefficients. Section 2 gives results for the out-of-sample analysis using the regression coefficients in Section 1. Section 3 gives a deeper analysis of negative log-likelihood statistics for the test and the training data sets for both Poisson and multi-linear regression. Section 4 conducts additional study into the findings presented in Section 1-3 of this chapter. The final section validates the computer code used throughout the study. See Appendix D for an executive summary of the result findings.

4.1 Model Evaluation Using In-Sample Training Data

The results in this section are from models built using the training data - data sets, $D_1, D_2, \ldots, D_{10}$. For both the multi-linear and Poisson regression, the best individual socio-economic factor to predict the number of recruits is QMA followed by the number of micro zip codes. However, the explanatory variable with the most predictive power is the number of recruiters, which is not a socio-economic factor but a measure of the resources the Army devotes to recruiting. The multi-linear regression models have the most predictive power. Version 16, the version with all five socio-economic factors plus a constant and the number of recruiters, has the most explanatory power for both Poisson and multi-linear regression models. Section 4.2 reviews how the models fit the final test data set and explain in detail how the models perform on data that was neither used for model selection nor model fitting.

4.1.1 Utilizing Negative Log-Likelihood – In Sample Data

Although Poisson regression and multi-linear regression are fundamentally different and built under very different assumptions, one compares Poisson regression to multi-linear regression through a negative log-likelihood statistic. The negative log-likelihood function evaluates the likelihood of seeing the observed values under the

conditions of the model. A smaller NLL value indicates a higher likelihood of the data under the model and means the model better fits the data.

In model versions 1-5, each socio-economic factor is evaluated individually. The model version with the smallest NLL for an individual socio-economic factor is model version 5, QMA. See Table 3 for a list of all the model versions and their likelihood statistics. Bold entries indicate which NLL statistic is lower for that model version. The MLR gives a smaller NLL for all 16 model versions.

Model Versions		Poisson - NLL	MLR - NLL
1	Unemployment Rate, Constant	4284	**3466**
2	Metro, Constant	4416	**3495**
3	Micro, Constant	4265	**3465**
4	Other, Constant	4314	**3476**
5	QMA, Constant	4158	**3435**
6	Recruiter, Constant	3539	**3256**
7	Unemployment Rate, Recruiter, Constant	3504	**3243**
8	Metro, Recruiter, Constant	3523	**3250**
9	Micro, Recruiter, Constant	3533	**3254**
10	Other, Recruiter, Constant	3533	**3254**
11	QMA, Recruiter, Constant	3534	**3254**
12	Unemployment Rate, Metro, Recruiter, Constant	3496	**3239**
13	Unemployment Rate, Micro, Recruiter, Constant	3500	**3242**
14	Unemployment Rate, Other, Recruiter, Constant	3502	**3243**
15	Unemployment Rate, QMA, Recruiter, Constant	3496	**3240**
16	Unemployment Rate, Metro, Micro, Other, QMA, Recruiter, Constant	3485	**3236**

Table 3. List of versions and the average out-of-sample negative log-likelihood statistic for all Poisson and MLR models. These results utilize 9,000 observations randomly selected and partitioned into 10 sets, $D_1, D_2, \ldots, D_{10}$. Recall Section 3.4 for a description of the cross-validation procedure, Section 3.3 for the calculation of a negative log-likelihood for Poisson regression, and Section 3.2.1, for the calculation of a negative log-likelihood for MLR. Bold entries indicate which NLL statistic is lower for that model version. The MLR gives a smaller NLL for all 16 models when comparing within each version.

In all model versions, the multi-linear regression is better than the Poisson regression models. In fact, the MLR models perform better than almost all of the Poisson regression models. The worst fitting MLR model (version 2) has a NLL value of 3495. This is bested by only one Poisson regression model (version 16) with a NLL value of 3485. Section 4.3 discusses an explanation for this.

4.1.2 Regression Coefficients for Poisson and Multi-Linear Regression Models

The regression coefficients for Poisson regression and MLR have different but meaningful interpretations. The interpretation for the Poisson regression coefficients is as follows: for a one unit change in the explanatory variable (x_{in}), the difference in the natural log of the expected value of the dependent variable (λ) is the regression coefficient (β_n), given the other explanatory variables in the model are held constant. Recall the form of a Poisson regression:

$$\ln(E(Y)) = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} .$$

More simply put, for a one unit change in the explanatory variable (x_{in}), the change in the expected value of the dependent variable ($\delta E(Y)$) is an exponential transformation of the regression coefficient (β_n).

$$\delta E(Y) = e^{\beta_n x_{in}}.$$

See Table 4 for the regression coefficients for all versions of Poisson regression. The description of the explanatory variable is respective to the x_i column heading. For example, in model version 1, the unemployment rate regression coefficient is in the x_1 column, the constant regression coefficient is in column x_2

Model Versions		x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	Unemployment Rate, Constant	0.05	2.96					
2	Metro, Constant	6.2E-4	3.29					
3	Micro, Constant	-3.7E-3	3.44					
4	Other, Constant	-1.7E-3	3.41					
5	QMA, Constant	3.0E-6	2.99					
6	Recruiter, Constant	0.03	2.38					
7	Unemployment Rate, Recruiter, Constant	0.02	0.03	2.23				
8	Metro, Recruiter, Constant	-8.0E-4	0.03	2.43				
9	Micro, Recruiter, Constant	-7.7E-4	0.03	2.43				
10	Other, Recruiter, Constant	-3.9E-4	0.03	2.42				
11	QMA, Recruiter, Constant	0E+0	0.03	2.39				
12	Unemployment Rate, Metro, Recruiter, Constant	0.02	-6.0E-4	0.03	2.28			
13	Unemployment Rate, Micro, Recruiter, Constant	0.02	-6.0E-4	0.03	2.27			
14	Unemployment Rate, Other, Recruiter, Constant	0.02	-2.1E-4	0.03	2.25			
15	Unemployment Rate, QMA, Recruiter, Constant	0.03	-1.0E-6	0.03	2.23			
16	Unemployment Rate, Metro, Micro, Other, QMA, Recruiter, Constant	0.02	-4.32E-4	-7.8E-4	-7.7E-5	-1.0E-6	0.03	2.33

Table 4. List of regression coefficients for Poisson regression models. These coefficients result from 9,000 randomly selected observations. See Section 3.4 for a description of modeling training.

The regression coefficients for MLR are listed in Table 5. The interpretation for these coefficients is more intuitive than for Poisson regression. It is as follows: for every unit change in the explanatory variable, the dependent variable changes by the regression coefficient provided that all other explanatory variables are held constant. For example, in version 12, for each additional recruiter, the model predicts an additional 0.85 recruits

while holding the unemployment rate, number of metro zip codes, and constant variables, steady. Some models did exhibit multi-coliniarity with the explanatory variables, meaning the coefficients we present may not be unique. However, these are the coefficients that produce the negative log likelihoods seen in Table 3. Although some multi-coliniarity is present, coefficients such as the ones listed are still useful as they present a method USAREC can apply in order to set recruiting goals.

Model Versions		x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	Unemployment Rate, Constant	1.45	17.12					
2	Metro, Constant	0.02	26.70					
3	Micro, Constant	-0.10	31.14					
4	Other, Constant	-0.04	30.11					
5	QMA, Constant	0.00	17.79					
6	Recruiter, Constant	0.86	0.83					
7	Unemployment Rate, Recruiter, Constant	0.74	0.82	-3.80				
8	Metro, Recruiter, Constant	-0.02	0.89	2.27				
9	Micro, Recruiter, Constant	-0.02	0.83	2.02				
10	Other, Recruiter, Constant	-0.01	0.84	1.72				
11	QMA, Recruiter, Constant	0.00	0.91	1.05				
12	Unemployment Rate, Metro, Recruiter, Constant	0.66	-0.02	0.85	-2.22			
13	Unemployment Rate, Micro, Recruiter, Constant	0.72	-0.01	0.80	-2.81			
14	Unemployment Rate, Other, Recruiter, Constant	0.72	0.00	0.81	-3.28			
15	Unemployment Rate, QMA, Recruiter, Constant	0.78	0.00	0.89	-3.75			
16	Unemployment Rate, Metro, Micro, Other, QMA, Recruiter, Constant	0.69	-0.01	-0.02	-1E-03	-2E-05	0.88	-0.99

Table 5. List of regression coefficients for MLR models. These coefficients result from 9,000 randomly selected observations. See Section 3.4 for a description of modeling training.

4.2 Model Evaluation Using Out-of-Sample Test Data

As described earlier, 1,323 randomly selected observations of the available data are set aside to use as a final test data set. The model selection is performed without including these observations. Therefore, the NLL calculated while using parameters set by the models built from the training data serve as an estimate of the true predictive power of the models. Again, of the models for individual socio-economic factors, QMA has the smallest NLL and thereby the greatest explanatory power followed by micro and the unemployment rate.

The calculation for the out-of-sample MLR NLL is slightly different than for the in-sample MLR NLL. When the in-sample NLL is calculated, the study captures the σ^2 value specific to the parameters of the trained model. For the MLR NLL on out-of sample data, the study utilizes the trained σ^2 to calculate NLL along with the regression coefficients of the trained model.

As with the models formed from the training data, MLR outperforms Poisson regression. See Table 6 for the negative log-likelihoods. On average, the NLL for MLR is 11% smaller than the NLL for Poisson regression. Moreover, the worst performing multi-linear regression model (NLL value of 5124) is better than the best performing Poisson regression model (NLL value of 5163). Bold entries indicate the smaller NLL statistics for that model version.

Model Versions		Poisson - NLL	MLR - NLL
1	Unemployment Rate, Constant	6369	**5124**
2	Metro, Constant	6532	**5159**
3	Micro, Constant	6316	**5116**
4	Other, Constant	6395	**5134**
5	QMA, Constant	6141	**5070**
6	Recruiter, Constant	5224	**4804**
7	Unemployment Rate, Recruiter, Constant	5195	**4794**
8	Metro, Recruiter, Constant	5207	**4797**
9	Micro, Recruiter, Constant	5220	**4803**
10	Other, Recruiter, Constant	5220	**4803**
11	QMA, Recruiter, Constant	5209	**4798**
12	Unemployment Rate, Metro, Recruiter, Constant	5185	**4789**
13	Unemployment Rate, Micro, Recruiter, Constant	5192	**4793**
14	Unemployment Rate, Other, Recruiter, Constant	5194	**4794**
15	Unemployment Rate, QMA, Recruiter, Constant	5176	**4786**
16	Unemployment Rate, Metro, Micro, Other, QMA, Recruiter, Constant	5163	**4781**

Table 6. List of NLL for all Poisson and MLR models evaluated using the final test data set. The final test data set consist of 1,323 randomly selected observations of the total data set. Bold entries indicate the smaller NLL statistics for that model version. Observe the worst performing MLR model has a smaller NLL statistic (5124) than the best performing Poisson regression model (5163).

4.3 Deeper Analysis of Negative Log-Likelihood Results

Using out of sample cross validation allows the study to gain additional insight on the models created. To calculate the NLL, the study uses ten-fold cross validation outlined in Section 3.4. As a result, each NLL calculation for the training data is an

average of 10 sets of 900 terms. The NLL calculations for models using the final test data set each have 1 set of 1,323 terms. The larger number of terms for the final test data set means a larger error value when compared to the training data error. However, the increase in the error value is in proportion to the ration of the training data set to the final test data set. In other words, 900 is approximately 67% of 1,323 and the NLL values for the training data is approximately 65% of the NLL values of the test data.

To highlight the models do not over-fit the data during the model selection step, we present a side-by-side listing of the NLLs for Poisson and MLR, test and training data sets respectively. This shows the ordering of models is consistent between the test and training data. See Table 7 (on next page) for the NLL values both the test and training data for Poisson and multi-linear regression.

Model Versions		Poisson (test)	Poisson (train)	MLR (test)	MLR (train)
1	Unemployment Rate, Constant	6369	4284	5124	3466
2	Metro, Constant	6532	4416	5159	3495
3	Micro, Constant	6316	4265	5116	3465
4	Other, Constant	6395	4314	5134	3476
5	QMA, Constant	6141	4158	5070	3435
6	Recruiter, Constant	5224	3539	4804	3256
7	Unemployment Rate, Recruiter, Constant	5195	3504	4794	3243
8	Metro, Recruiter, Constant	5207	3523	4797	3250
9	Micro, Recruiter, Constant	5220	3533	4803	3254
10	Other, Recruiter, Constant	5220	3533	4803	3254
11	QMA, Recruiter, Constant	5209	3534	4798	3254
12	Unemployment Rate, Metro, Recruiter, Constant	5185	3496	4789	3239
13	Unemployment Rate, Micro, Recruiter, Constant	5192	3500	4793	3242
14	Unemployment Rate, Other, Recruiter, Constant	5194	3502	4794	3243
15	Unemployment Rate, QMA, Recruiter, Constant	5176	3496	4786	3240
16	Unemployment Rate, Metro, Micro, Other, QMA, Recruiter, Constant	5163	3485	4781	3236

Table 7. Side by side comparison of the NLL statistic for test and training data for Poisson and multi-linear regression respectively. There are 1,323 observations in the test results and a total of 9,000 observations in the training results. The training and test data sets are partitioned by randomly indexing the total set of observations into 11 bins. The eleventh bin containing 1,323 observations is the test data set. The training data set is split into 10 bins of 900 observations each, and the NLLs presented above are averages over leave-one-bin-out evaluations of NLL for these 10 bins. This demonstrates the model is not over-fitting the data during the model selection step.

As mentioned in Sections 4.1.1 and 4.2, the NLLs for the MLR models are consistently better than the Poisson regression models. One explanation is simply that MLR is a better tool than Poisson regression for modeling this data. Another explanation

is that the NLLs for MLR are better than the NLLs for Poisson regression is because the MLR NLL has an additional parameter, σ^2, which helps to minimize the overall statistic. However, in the test set results presented in Table 7, all MLR parameters, including σ^2 were derived from the training set. This provides strong evidence that the superior MLR results are not due to over-fitting the data, but better modeling capability.

4.4 Mini-Study of Main Results – Predicting Recruits per Recruiter

One may argue the number of recruits is strongly linearly dependent on the number of recruiters, and that is why multi-linear regression models outperform Poisson regression models. In addition, number of recruiters is not a true socio-economic factor. To address these issues, we perform an additional analysis that transforms the dependent variable *y* from the number of recruits to the *recruiter rate*. The recruiter rate is the number of recruits per recruiter. Furthermore, the study modifies the model space to reflect the transformation of *y* and includes QMA as an explanatory variable in more models due to the initial evidence that QMA is the factor with most explanatory power. Next, the study re-calculates the NLL statistics for all of the training and test data sets. See Table 8 for a list of new model versions. Let *MS* stand for mini-study.

Version	Explanatory Variable					
1_{MS}	x_{i1} = Unemployment Rate					$\beta_0 x_{i0}$ = constant
2_{MS}	x_{i1} = Metro					$\beta_0 x_{i0}$ = constant
3_{MS}	x_{i1} = Micro					$\beta_0 x_{i0}$ = constant
4_{MS}	x_{i1} = Other					$\beta_0 x_{i0}$ = constant
5_{MS}	x_{i1} = QMA					$\beta_0 x_{i0}$ = constant
6_{MS}	x_{i1} = QMA		x_{i2} = Metro			$\beta_0 x_{i0}$ = constant
7_{MS}	x_{i1} = QMA		x_{i2} = Micro			$\beta_0 x_{i0}$ = constant
8_{MS}	x_{i1} = QMA		x_{i2} = Other			$\beta_0 x_{i0}$ = constant
9_{MS}	x_{i1} = QMA		x_{i2} = Unemployment Rate			$\beta_0 x_{i0}$ = constant
10_{MS}	x_{i1} = Unemployment Rate	x_{i2} = Metro	x_{i3} = Micro	x_{i4} = Other	x_{i5} = QMA	$\beta_0 x_{i0}$ = constant

Table 8. Mini-study model versions and explanatory variables.

4.4.1 Negative Log-Likelihood Analysis - Recruiter Rate

The table below presents NLLs for both the training and test data sets for MLR and Poisson regression. The NLLs are orders of magnitude better for the mini-study than for the main-study because the constant is an excellent predictor. By using the constant as the sole predictor in both Poisson regression and MLR, the study observes how much the additional parameters are improving the fit of the model. The MLR show measurable improvement whereas the Poisson regression does not. See Table 9 for a list of NLL results. Remember the MLR NLL calculation for out-of-sample data uses the σ^2 values from the in-sample, trained models.

Version	MLR (training)	MLR (test)	Poisson (training)	Poisson (test)
constant	124	200	877	1292
1_{MS}	112	192	876	1291
2_{MS}	119	191	877	1291
3_{MS}	122	199	877	1292
4_{MS}	122	199	877	1292
5_{MS}	123	194	877	1291
6_{MS}	119	189	877	1291
7_{MS}	118	187	877	1291
8_{MS}	119	188	877	1291
9_{MS}	108	181	876	1290
10_{MS}	103	174	876	1289

Table 9. NLL values for both the training and test data sets for both MLR and Poisson regression. There are 1,323 observation in the test results and a total of 9,000 observations in the training results. The training and test data sets are partitioned by randomly indexing the total set of observations into 11 bins. The eleventh bin contains 1,323 observations and is the test data set. Bins 1-10 each contained 900 observations, their union is the training data set. See Section 3.4 to review of the cross-validation procedure. By using the constant as the sole predictor in both Poisson regression and MLR, the study observes how much the additional parameters are improving the fit of the model. The MLR show measurable improvement whereas the Poisson regression does not.

Although the MLR NLL has an additional parameter,σ^2, the difference between MLR and Poisson regression is drastic. On average, the NLL for Poisson regression is 7 times worse than the NLL for the MLR regression. Further, the performance of MLR on the test sets, where both the coefficients and ,σ^2 are determined from the training data provides evidence that MLR provides more robust modeling of this data. Another explanation for the degraded performance of the Poisson regression in relation to the

MLR regression in the mini-study is that the dependent variable no longer fits the definition of count data.

4.4.2 Mini-Study – Recruiter Rate Result

The essential result of this mini-study into predicting the recruiter rate is that average recruiter rate is a great predictor for the future recruiter rate. Socio-economic factors do add additional explanatory power, particularly for the MLR models. The mini-study into predicting recruiter rate provides further evidence that MLR models are a better fit for the recruiting data. Without this study, one may argue that numbers of recruits has a strongly linear dependence on the number of recruiters but the arrivals per recruiter are Poisson in the other explanatory variables. The mini-study provides evidence to the contrary. Chapter 5 addresses the results further.

4.5 Code Check

These results presented in this chapter are counterintuitive because the number of recruits at the recruiting company level fits the classic definition of count data - data that can only take non-negative integer values. To check the validity of the computer code, the study tests it on a data set specifically designed to produce better results for Poisson regression than MLR. The test steps are as follows:

Step 1: Let t_{i0} and t_{i1} be explanatory variables. For $i = 1, 2, \dots, 30$, let $t_{i0} = 1$ and t_{i1} be a random number between 0 and 1.

Step 2: Let j_0 be the regression coefficient for t_{i0}. Set j_0 equal to 3. Let j_1 be the regression coefficient for t_{i1}. Set j_1 equal to 2.

Step 3: Let v_i be the dependent variable, where $v_i = e^{t_{i0}j_{i0} + t_{i1}j_{i1}}$

Step 4: Utilizing the same code as the rest of the study, calculate the NLL for MLR and Poisson regression on this generated dataset. Give the regression coefficients for both Poisson and MLR.

The results of this check give the NLL for Poisson regression as 89 and the NLL for MLR as 108. The Poisson regression also finds the exact values provided for j_0 and j_1. Recall in Section 4.1.1, the study cautions against expecting a 1 for the R^2 of a Poisson regression. In fact, the R^2 value for this code check is only 0.76 despite finding the exact regression coefficients. The results of this code check allow additional confidence in the results.

Chapter 5. Discussion and Future Work

The results of Chapter 4 are somewhat counter-intuitive because, outwardly, the data appears to be count data for which Poisson regression is suited. However, when holding model version consistent, be it the main or mini-study, multi-linear regression outperforms Poisson regression. Although multi-linear regression is the better performer, the results of the main-study – and particularly the mini-study – support the notion from the related works that unemployment rate is a factor driving the number of recruits. While recruiters drive recruitment, both the main and mini-studies show socio-economic factors do add significant, validated, out-of sample predictive capability – particularly for multi-linear regression. These results warrant a closer look at the data set and the current method of Army recruiting procedures.

As covered in Chapter 3, the data consists of 10,323 observations encompassing four sequential fiscal years' worth of recruiting data for 250 recruiting companies. The limited years of the data set may provide a level of consistency in the socio-economic factors. Increasing the amount of historical data to 10 years or more allows for a greater range in the values of the socio-economic factors. This could, in turn, give a clearly picture of how that factor affects the number of recruits.

Another possible issue with the data set is the accuracy of the information. A small investigation of the data reveals potential issues. For example, the number of metro, micro, and other zip codes is problematic. Recall that metro means a zip code with a population of 50,000 or more; micro has a population of 10,000 to 50,000; and other has a population of 10,000 or less. Using these parameters, the study calculates a very conservative population estimate for the United States by setting the population size of a metro, micro, and other zip codes to 50,000, 10,000, and 100 respectively. Summing the population estimate across all recruiting companies for January in a fiscal year 2012 should yield a result close to the current total population of the United States. The current population of the United States is 318.9 million (US Census Bureau, 2014). Using the data provided and conservatively setting their values, the study finds a

population estimate of 1,222,910,200. Unless the inaccuracy in the data is consistent throughout the entire data set, it will detrimentally affect results.

The current method for recruiting goals relies on a weighted average of pervious recruits (See Appendix B). Utilizing this method makes the data somewhat circular. Outside forces, other than the level of troop demand as set by the executive branch of the government, are not factored in the current procedure. Recruiters are pressured to achieve what was done previously. In other words, perhaps future study requires an additional explanatory variable measuring the amount of pressure recruiters receive from the Army to meet quota.

The study only includes five socio-economic factors. In the scheme of possible socio-economic factors, this is a very small amount. Other socio-economic factors, such as the vast array of factors collected by the American Community Survey (American Community Survey, 2015), require additional exploration. However, adding these factors requires a sufficiently fine-grained estimate, both geographically and in time, as this survey is conducted every five years. Other avenues of socio-economic factor investigation include PRISM segments. While the study creates models without using PRISM segments, the socio-economic factors driving that segmentation deserve study. One could apply Poisson regression methodology within each PRISM segment at the company level. Performing regression studies within each PRISM segment likely requires additional statistical care, since many segments have low population numbers and the data would likely exhibit a large number of zero values for the dependent variable.

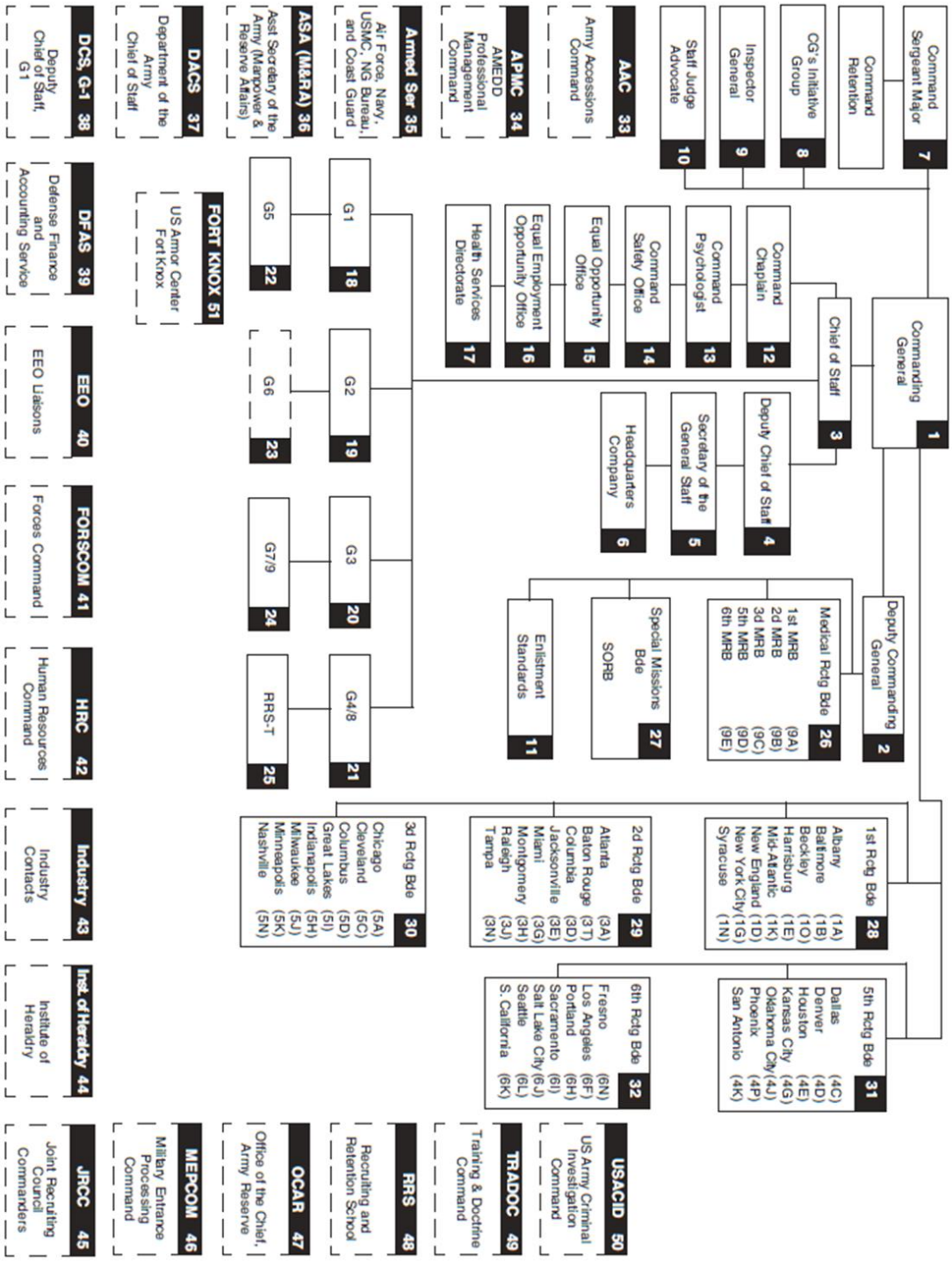
One of the study's objectives is to replace the Army's current multi-step process for calculating recruiting goals, SAMA, with a single calculation at the recruiting station level. While the study does not definitively show this is reasonable, it does present another area of future work. SAMA relies on a four year historical weighted average. As the study only has four years' worth of data, it cannot currently compare SAMA to the models in the study. Furthermore, SAMA is not a probabilistic model like the models in

the study. Future work could include modeling SAMA as a time-series exponential smoothing model and then comparing its NLL to the models in this study.

Chapter 6. Conclusion

The goals of this study are two-fold: 1) find a simple and accurate prediction method for recruiting capacity and 2) identify the most significant socio-economic factors in determining recruiting capacity. The report achieves these goals through building two types of models, Poisson regression and multi-linear regression models. Despite the difficulties outlined in Chapter 5 and only considering five socio-economic factors, the multi-linear regression models clearly demonstrate consistent explanatory power, more so than Poisson regression. As such, future modeling of recruits should utilize multi-linear regression over Poisson regression. Moreover, multi-linear regression models should include the number of recruiters as an explanatory variable when recruits is the dependent variable. The most predictive socio-economic factors available in this study are QMA and unemployment rate when modeling the number of recruits and recruiter rate respectively. This supports the notion from the related works that the unemployment rate is a major factor in recruiting ability.

APPENDIX A.


Command Sergeant Major 7
Command Retention
CG's Initiative Group 8
Inspector General 9
Staff Judge Advocate 10
Commanding General 1
Chief of Staff 3
Deputy Commanding General 2
Command Chaplain 12
Command Psychologist 13
Command Safety Office 14
Equal Opportunity Office 15
Equal Employment Opportunity Office 16
Health Services Directorate 17
Deputy Chief of Staff 4
Secretary of the General Staff 5
Headquarters Company 6
Medical Rctg Bde 26
1st MRB (9A)
2d MRB (9B)
3d MRB (9C)
5th MRB (9D)
6th MRB (9E)
Special Missions Bde 27
SORB
Enlistment Standards 11
1st Rctg Bde 28
Albany (1A)
Baltimore (1B)
Beckley (1O)
Harrisburg (1E)
Mid-Atlantic (1K)
New England (1D)
New York City (1G)
Syracuse (1N)
5th Rctg Bde 31
Dallas (4C)
Denver (4D)
Houston (4E)
Kansas City (4G)
Oklahoma City (4J)
Phoenix (4P)
San Antonio (4K)
2d Rctg Bde 29
Atlanta (3A)
Baton Rouge (3T)
Columbia (3D)
Jacksonville (3E)
Miami (3G)
Montgomery (3H)
Raleigh (3J)
Tampa (3N)
6th Rctg Bde 32
Fresno (6N)
Los Angeles (6F)
Portland (6H)
Sacramento (6I)
Salt Lake City (6J)
Seattle (6L)
S. California (6K)
3d Rctg Bde 30
Chicago (5A)
Cleveland (5C)
Columbus (5D)
Great Lakes (5I)
Indianapolis (5H)
Milwaukee (5J)
Minneapolis (5K)
Nashville (5N)
AAC 33
Army Accessions Command
APMC 34
AMEDD Professional Management Command
Armed Ser 35
Air Force, Navy, USMC, NG Bureau, and Coast Guard
ASA (M&RA) 36
Asst Secretary of the Army (Manpower & Reserve Affairs)
DACS 37
Department of the Army Chief of Staff
DCS, G-1 38
Deputy Chief of Staff, G1
DFAS 39
Defense Finance and Accounting Service
EEO 40
EEO Liaisons
FORSCOM 41
Forces Command
HRC 42
Human Resources Command
Industry 43
Industry Contacts
Inst. of Heraldry 44
Institute of Heraldry
JRCC 45
Joint Recruiting Council Commanders
MEPCOM 46
Military Entrance Processing Command
OCAR 47
Office of the Chief, Army Reserve
RRS 48
Recruiting and Retention School
TRADOC 49
Training & Doctrine Command
USACID 50
US Army Criminal Investigation Command
FORT KNOX 51
US Armor Center Fort Knox
G1 18
G2 19
G3 20
G4/8 21
G5 22
G6 23
G7/9 24
RRS-T 25

APPENDIX B.

Currently, the Army computes the number of potential recruits through a multi-step process that is essentially a weighted average of recruiting data from the last four years. The most recent year is weighted heaviest at 0.4, with each subsequent year weighted one tenth less. For example, when determining the goal for target year 2015, the recruiting data in 2014, 2013, 2012, and 2011 are weighted 0.4, 0.3, 0.2, and 0.1 respectively. The following details how the current recruiting potential is calculated:

Step 1: Partition zip codes by demography. Each zip code is divided into 39 *tactical segments* (TS). Tactical segments are determined based on three demographic characteristics, race, age, and gender. For example, tactical segment 4, in any zip code, consists of all the Caucasian males, age 16-19.

Step 2: Determine the best potential penetration rate, PR_{best}. The *penetration rate* (PR) is the percentage of the population successfully recruited.

1. Let $4YrWtCalc_{TS}$, where *TS* is a particular tactical segment, be the weighted average of the last four years of recruiting contracts produced. Let Y_i, where *i* is the number of years prior to the target year, be the number of contracts produced in that tactical segment.
    a. $4YrWtCalc_{TS} = (Y_{1,TS} * 0.4) + (Y_{2,TS} * 0.3) + (Y_{3,TS} * 0.2) + (Y_{4,TS} * 0.1)$
2. Let PR_{TS} be the penetration rate for a particular tactical segment and Pop_{TS} be the total population in that particular tactical segment.
    a. $PR_{TS} = 4YrWtCalc_{TS} / Pop_{TS}$
3. Let PR_{zip} be the penetration rate for a particular zip code. Sum the PR_{TS} for all tactical segments in that zip code.
    a. $PR_{zip} = \sum_{TS=1}^{39} PR_{TS}$
4. Let PR_{ST} be the penetration rate for a recruiting station. Let *st_zips* be the set of all zip codes belonging to a particular recruiting station

a. $PR_{ST} = \sum_{st\_zips} PR_{zip}$

5. Let PR_{CO} be the penetration rate for a recruiting company. Let *co_sts* be the set of all recruiting stations belonging to a particular recruiting company.
   a. $PR_{CO} = \sum_{co\_sts} PR_{ST}$
6. Determine the best potential Penetration Rate.
   a. $PR_{best} = \max(PR_{ST}, PR_{CO})$

Step 3: Determine the *Army volume potential* (AVP) for each zip code, AVP_{zip}.

1. Let $Potential\_Production_{TS}$ be the Army potential production by tactical segment.
   a. $Potential\_Production_{TS} = PR_{best} * Pop_{TS}$
2. Let $Potential_Production_{zip}$be the Army potential production by zip code.
   a. $Potential\_Production_{zip} = \sum_{TS=1}^{39} Potential\_Production_{TS}$
3. Sometimes, recruits do not fall into a specified tactical segment for whatever reason, however, these numbers must still be captured when determining AVP. Let NC_{zip} be the four year weighted average number of contracts in a particular zip code. Let $X_{i,zip}$, where *i* is the number of years prior to the target year and *zip* is a particular zip code, be the number of contracts produced that do not fall into a specific tactical segment.
   a. NC_{zip} = ($X_{1,zip}$ * 0.4) + ($X_{2,zip}$ * 0.3) + ($X_{3,zip}$ * 0.2) + ($X_{4,zip}$ * 0.1)
4. Calculate AVP_{zip}.
   a. $AVP_{zip} = Potential\_Production_{zip} + NC_{zip}$

APPENDIX C. R^2 Analysis

In both Poisson and MLR, the models with the most explanatory power include all five socio-economic factors plus a constant and the number of recruiters. Although the bulk of the R^2 value is explained by the number of recruiters, the socio-economic factors do add additional explanatory power. The study concludes there are likely other socio-economic factors capable of adding additional explanation for the number of recruits. This appendix reviews R^2 results for in-sample data and out-of-sample data followed by further analysis of both. It concludes with R^2 analysis of the mini-study models.

IN SAMPLE DATA

The study utilizes five different socio-economic factors - unemployment rate, the number of metro, micro, and other zip codes, and the population of qualified military aged civilians, plus a measurement of the Army's efforts, the number of recruiters – as the explanatory variables available to build each regression model. Each explanatory variable is examined individually then later in combination with others. The R^2 values for the models that use a single socio-economic factor allow for the identification of the factor with the most explanatory power.

In multi-linear regression (MLR), the R^2 value calculation compares the regression sum of squared error (SSE) to the SSE of the best constant predictor. In this study, the best constant predictor is the mean of all the observed recruits. The closer R^2 is to one, the better the model fits the data. The calculation is as follows:

$$R^2 = 1 - \frac{\sum_i \left(y_i - y(x_i)_{predict}\right)^2}{\sum_i (y_i - y_{mean})^2} = 1 - \frac{SSE_{regression}}{SSE_{constant}}.$$

where $y(x_i)_{predict}$ is the prediction given inputs x_i and y_{mean} is the best constant predictor.

The R^2 in a Poisson regression model is not the same R^2 in multi-linear regression. Instead, a $psuedo\ R^2$ value calculation utilizes the same principle of

examining the ratio between the regression model error against the best constant predictor. More specifically, this means comparing the negative log-likelihood (NLL) of the regression model to the negative log-likelihood of a model that assumes a constant parameter. To calculate $NLL_{constant}$, y_i is regressed against the constant 1. The $pseudo\ R^2$ value is defined as:

$$psuedo\ R^2 = 1 - \frac{NLL_{regression}}{NLL_{constant}}.$$

Although the principle for calculating R^2 for MLR and Poisson regression is the same, it is a mistake to compare the R^2 values for a MLR to the *pseudo* R^2 of a Poisson. These values are only useful for comparing one model version to another within the same regression type. It is important to note that the R^2 value for a Poisson model can never be 1, like it can for a multi-linear regression, because the negative log-likelihood for a Poisson regression will never go to 0. This is another reason why comparing the R^2 value for a MLR to a Poisson is a mistake. Another potential issue with using an R^2 value to evaluate a model is that if the model over-fits the training data, R^2 can actually be negative. In other words, a negative R^2 means that the best constant predictor is a better than using a model with an explanatory variable. See Table 10 for a list of the model versions and the associated R^2 values. Bolded entries indicate the models with the best performing socio-economic factors and the model with the highest overall R^2 value.

Model Versions		Poisson - R^2	MLR - R^2
1	**Unemployment Rate, Constant**	**0.033**	**0.068**
2	Metro, Constant	0.002	0.005
3	**Micro, Constant**	**0.037**	**0.070**
4	Other, Constant	0.026	0.046
5	**QMA, Constant**	**0.061**	**0.128**
6	Recruiter, Constant	0.201	0.415
7	Unemployment Rate, Recruiter, Constant	0.209	0.432
8	Metro, Recruiter, Constant	0.204	0.424
9	Micro, Recruiter, Constant	0.202	0.417
10	Other, Recruiter, Constant	0.202	0.417
11	QMA, Recruiter, Constant	0.200	0.418
12	Unemployment Rate, Metro, Recruiter, Constant	0.202	0.437
13	Unemployment Rate, Micro, Recruiter, Constant	0.211	0.433
14	Unemployment Rate, Other, Recruiter, Constant	0.209	0.432
15	Unemployment Rate, QMA, Recruiter, Constant	0.209	0.436
16	**Unemployment Rate, Metro, Micro, Other, QMA, Recruiter, Constant**	**0.210**	**0.441**

Table 10. List of model versions and R^2values for Poisson regression and MLR models. These results utilize 9,000 observations randomly selected and partitioned into 10 sets, $D_1, D_2, \ldots, D_{10}$. All 9,000 observations are used as per the model training described in Section 3.4. Bolded entries indicate the models with the best performing socio-economic factors and the model with the highest overall R^2 value.

OUT OF SAMPLE DATA

As described earlier, 1,323 randomly selected observations of the available data are set aside to use as a final test data set. The model selection is performed without including these observations. Therefore, the R^2values calculated while using the parameters set by the models built from the training data serve as an estimate of the true predictive power of the models. Again, QMA has the most explanatory power followed by micro and the unemployment rate. See Table 11 for the R^2values. Bolded entries indicate the models with the best performing socio-economic factors and the model with the highest overall R^2 value.

Model Versions		Poisson - R^2	MLR - R^2
1	**Unemployment Rate, Constant**	**0.057**	**0.056**
2	Metro, Constant	0.033	0.005
3	**Micro, Constant**	**0.065**	**0.067**
4	Other, Constant	0.054	0.042
5	**QMA, Constant**	**0.091**	**0.130**
6	Recruiter, Constant	0.227	0.418
7	Unemployment Rate, Recruiter, Constant	0.231	0.427
8	Metro, Recruiter, Constant	0.229	0.425
9	Micro, Recruiter, Constant	0.227	0.419
10	Other, Recruiter, Constant	0.227	0.419
11	QMA, Recruiter, Constant	0.229	0.424
12	Unemployment Rate, Metro, Recruiter, Constant	0.233	0.431
13	Unemployment Rate, Micro, Recruiter, Constant	0.232	0.428
14	Unemployment Rate, Other, Recruiter, Constant	0.231	0.427
15	Unemployment Rate, QMA, Recruiter, Constant	0.234	0.434
16	**Unemployment Rate, Metro, Micro, Other, QMA, Recruiter, Constant**	**0.236**	**0.438**

Table 11. R^2 values for Poisson and MLR models evaluated using the final test data set. The final test data set consist of 1,323 randomly selected observations of the total data set. Bolded entries indicate the models with the best performing socio-economic factors and the model with the highest overall R^2 value.

FURTHER ANALYSIS

The R^2 values for the Poisson regression and MLR models have an on average difference of 2% difference between the training and test data sets. The study expects that all the training data R^2 values to be higher than the test data, however, when the test data $R^2$is higher, it is only marginally so. See Table 12 for a comparison of R^2 values for both test and training data sets for Poisson and MLR. A green highlighted cell indicates the higher R^2 within that type of regression.

Model Versions		Poisson (test)	Poisson (train)	MLR (test)	MLR (train)
1	Unemployment Rate, Constant	0.027	0.033	0.056	0.068
2	Metro, Constant	0.002	0.002	0.005	0.005
3	Micro, Constant	0.035	0.037	0.067	0.070
4	Other, Constant	0.023	0.026	0.042	0.046
5	QMA, Constant	0.062	0.061	0.130	0.128
6	Recruiter, Constant	0.202	0.201	0.418	0.415
7	Unemployment Rate, Recruiter, Constant	0.207	0.209	0.427	0.432
8	Metro, Recruiter, Constant	0.205	0.204	0.425	0.424
9	Micro, Recruiter, Constant	0.203	0.202	0.419	0.417
10	Other, Recruiter, Constant	0.203	0.202	0.419	0.417
11	QMA, Recruiter, Constant	0.204	0.200	0.424	0.418
12	Unemployment Rate, Metro, Recruiter, Constant	0.208	0.202	0.431	0.437
13	Unemployment Rate, Micro, Recruiter, Constant	0.207	0.211	0.428	0.433
14	Unemployment Rate, Other, Recruiter, Constant	0.207	0.209	0.427	0.432
15	Unemployment Rate, QMA, Recruiter, Constant	0.209	0.209	0.434	0.436
16	Unemployment Rate, Metro, Micro, Other, QMA, Recruiter, Constant	0.212	0.210	0.438	0.441

Table 12. R^2 values for both test and training data sets for Poisson and MLR. There are 1,323 observation in the final test results and a total of 9,000 observations in the training results. The training and test data sets were partitioned by randomly indexing the total set of observations into 11 bins. The eleventh bin contains 1,323 observations and is the test data set. Bins 1-10 each contained 900 observations. The union of bins 1-10 is the training data set.

MINI-STUDY RECRUITER RATE RESULTS

The R^2results are somewhat surprising. In general, the R^2 values overall are much worse for MLR and Poisson regression than their main-study counterparts. This is because the constant is a good predictor for the recruiter rate. The socio-economic factors do add explanatory power but not nearly as much when compared to the main-study. Now, the leading socio-economic factor is the unemployment rate followed by the

number of metro zip codes. QMA, the previous leader, is last. The R^2 for the mini-study full model version, 10_{MS}, is surprisingly low. See Table 13 for a list of the R^2 values for both the training and test data sets for both MLR and Poisson regression.

Version	Socio-Economic Factors	MLR R^2 train	MLR R^2 test	Poisson R^2 train	Poisson R^2 test
1_{MS}	Unemployment Rate	0.027	0.012	1.2E-03	5.1E-04
2_{MS}	Metro	0.012	0.013	5.5E-04	5.9E-04
3_{MS}	Micro	0.005	0.001	2.2E-04	1.8E-05
4_{MS}	Other	0.005	0.001	2.3E-04	4.1E-05
5_{MS}	QMA	0.003	0.009	1.4E-04	3.8E-04
6_{MS}	QMA, Metro	0.012	0.015	5.6E-04	6.8E-04
7_{MS}	QMA, Micro	0.013	0.018	6.1E-04	8.1E-04
8_{MS}	QMA, Other	0.012	0.017	5.7E-04	7.5E-04
9_{MS}	QMA, Unemployment Rate	0.035	0.028	1.5E-03	1.2E-03
10_{MS}	Unemployment, Metro, Micro, Other, QMA	0.046	0.039	2.1E-03	1.8E-03

Table 13. R^2 values for both the training and test data sets for both MLR and Poisson regression.

There are 1,323 observation in the test results and a total of 9,000 observations in the training results. The training and test data sets were partitioned by randomly indexing the total set of observations into 11 bins. The eleventh bin contains 1,323 observations and is the test data set. Bins 1-10 each contained 900 observations. Their union is the training data set. The constant is the best predictor for the future recruiter rate and unemployment rate adds the most explanatory power.

APPENDIX D. Executive Summary

Utilizing Socio-Economic Factors to Evaluate Recruiting Potential for a US Army Recruiting Company

Sandra Young Jackson, M.S.E.
The University of Texas at Austin, 2015
SUPERVISOR: Nedialko B. Dimitrov
Co-SUPERVISOR: Jonathan K. Alt

In order to maintain military strength, the United States Army is consistently challenged with recruiting new soldiers. Currently the Army evaluates its recruiting capacity by calculating a weighted average of the previous four years of recruiting data. This report analyzes an alternative, statistical approach to computing recruiting capacity. Specifically, the study analyzes the effectiveness of multi-linear regression and Poisson regression models to compute recruiting capacity. The statistical analysis for these models is based on United States Army Recruiting Command data with 10,323 observations, encompassing four years of recruiting from 2011-2014. The data describes recruiting performance for each recruiting company, for each month, along with several other factors such as the number of recruiters in the company, the unemployment rate of the target region, and demographic descriptions of the target region.

The study analyzes two separate regression problems: predicting recruits, and predicting recruiter rates. For each of these problems the study constructs both multi-linear regressions and Poisson regressions, based on different subsets of explanatory variables and evaluates model performance on out-of-sample data. Out-of-sample evaluation increases confidence in statistical models because it demonstrates a level of performance on data that was not used to create the model.

Surprisingly, even though essentially all previous literature on recruiting suggests Poisson regression to model recruiting arrival rates, we show strong empirical evidence that multi-linear regression is a better modeling tool than Poisson regression for the recruiting data. On out-of-sample tests involving 32 competing models, the negative

log-likelihood for the multi-linear regression models is, on average over all the models, 11% smaller than the corresponding Poisson regression model. On out-of-sample tests involving an additional 20 models, the negative log-likelihood for the multi-linear regression is on average 85% smaller than the corresponding Poisson regression.

When the number of recruits is the dependent variable, for both the multi-linear and Poisson regression models, the best individual socio-economic factor to predict the number of recruits is the number of qualified military aged persons, followed by the number of micro zip codes. However, the explanatory variable with the most predictive power is the number of recruiters, which is not a socio-economic factor but a measure of the resources the Army devotes to recruiting. The multi-linear regression models have the most predictive power. A multi-linear regression model that includes the number of recruiters and five socio-economic factors has the most explanatory power.

When recruiter rate is the dependent variable, surprisingly, a constant is a great predictive model. Socio-economic factors, specifically the unemployment rate, do add additional explanatory power - particularly for the multi-linear regression models. The statistical analysis of recruiter rate suggests there is great potential for recruiting capacity because socio-economic factors do not limit the number of recruits. In other words, the results suggest that if the Army wants to increase recruits, one additional recruiter results in an additional 0.89 recruits.

Future work should include increasing years of historical data to compensate for possible homogeny in the time period of this study's data set. Furthermore, the study only includes five socio-economic factors. Other socio-economic factors, such as the vast array of factors collected by the American Community Survey (American Community Survey, 2015), require additional exploration. Another avenue of future study is to potentially apply regression models to recruiting within different PRISM segments, a system that demographically splits the population.

References

[1] Clingan, L. and Stokan, M. (September 2009). Segmentation Analysis Market Assessment [PDF document]. US Army Recruiting Command, G-2.

[2] Evans, M. and Powell, R. (July 2014). Nobel Index Technical Report Version 1. Navy Recruiting Command, Strategic Plans, Research, and Analysis. Millington, TN.

[3] Foti, S. G. (1978). *The importance of socio-economic factors in recruiting and sustaining the all-volunteer force* (Doctoral dissertation, Monterey, California. Naval Postgraduate School). [Online]. Available: http://www.dtic.mil/get-tr-doc/pdf?AD=ADA053876.

[4] Murray, M. P., & McDonald, L. L. (1998). *Recent recruiting trends and their implications for models of enlistment supply*. Rand national defense research inst santa monica ca. [Online]. Available: http://www.dtic.mil/get-tr-doc/pdf?AD=ADA360747.

[5] Pinelis, Y. K., Schmitz, E. J., Miller, Z. T., Rebhan, E. M., & Schmitz, E. J. (2011). An analysis of Navy recruiting goal allocation models.

[6] Quester, G. H. (2005). Demographic trends and military recruitment: Surprising possibilities. *Parameters*, *35*(1), 27-40.

[7] Schwartz, G. S. (March 1993). *Realigning the US Navy Recruiting Command* (Master Thesis, Monterey, California. Naval Postgraduate School). [Online]. Available: http://www.dtic.mil/get-tr-doc/pdf?AD=ADA265229.

[8] Stokan, M. (April 2014). Segmentation for Brigade & Battalion S2 Training [PDF document]. US Army Recruiting Command, G-2.

[9] United States Army Recruiting Command Headquarters. (September 2009). *USAREC Manual 3-0*. [Online]. Available: www.usarec.army.mil/im/formpub/REC_PUBS/man3_0.pdf

[10] US Army Recruiting Command. (August 2014). *USAREC About Us*. [Online]. Available: http://www.usarec.army.mil/aboutus.html.

[11] US Army Recruiting Command. (June 2014). *USAREC Electronic Publications*. [Online]. Available: http://www.usarec.army.mil/im/formpub/Pubs.htm.

[12] US Army Recruiting Command (May 2013). *USAREC May 2013 Talking Points*. [Online]. Available: http://www.usarec.army.mil/hq/apa/download/May2013talking-points.pdf.

[13] US Census Bureau (April 2014). *American Community Survey*. [Online]. Available: http://www.census.gov/acs/www/.

[14] US Census Bureau (July 2014). *US and World Population Clock.* [Online]. Available: http://www.census.gov/popclock/.

[15] Williams, T. (December 2014). *Understanding Factors Influencing Navy Recruiting Production.* (Master Thesis, Monterey, California. Naval Postgraduate School).